

Linear Batch Codes

Helger Lipmaa and Vitaly Skachek¹

Institute of Computer Science
University of Tartu

J. Liivi 2, Tartu 50409, Estonia

Email: {helger.lipmaa, vitaly.skachek} @ut.ee

Abstract—In an application, where a client wants to obtain many elements from a large database, it is often desirable to have some load balancing. Batch codes (introduced by Ishai *et al.* in STOC 2004) make it possible to do exactly that: the large database is divided between many servers, so that the client has to only make a small number of queries to every server to obtain sufficient information to reconstruct all desired elements. Other important parameters of the batch codes are total storage and the number of servers. Batch codes also have applications in cryptography (namely, in the construction of multi-query computationally-private information retrieval protocols).

In this work, we initiate the study of *linear* batch codes. These codes, in particular, are of potential use in distributed storage systems. We show that a generator matrix of a binary linear batch code is also a generator matrix of classical binary linear error-correcting code. This immediately yields that a variety of upper bounds, which were developed for error-correcting codes, are applicable also to binary linear batch codes. We also propose new methods to construct large linear batch codes from the smaller ones.

Index Terms—Batch codes, error-correcting codes, computationally-private information retrieval, load balancing, distributed storage.

I. INTRODUCTION

Consider the scenario where a client wants to retrieve many (say m) elements from an n element database, stored by a storage provider. Accessing a single server by all clients simultaneously can create serious performance problems. A simple solution is to duplicate the whole database between some M servers, so that the client can query approximately m/M elements from every server. However, that solution is very costly storage-wise, since all servers together have then to store $N = Mn$ database elements.

The things get even more problematic in the cryptographic scenario. In an m -out-of- n CIPR (computationally-private information retrieval [7]), the client wants to retrieve m elements from an n element database without the storage provider getting to know which elements were retrieved. An additional problem in this case is the storage provider's computational complexity that is $\Theta(n)$ per query in almost all known 1-out-of- n CIPR protocols. (The only exception is [8], where the per-query computational complexity is $O(n/\log n)$.) Just

performing m instances of an 1-out-of- n CIPR protocol would result in a highly prohibitive computational complexity.

To tackle both mentioned problems, Ishai *et al.* [6] proposed to use *batch codes*. More precisely, let Σ be a finite alphabet. In an $(n, N, m, M, T)_\Sigma$ batch code, a database f of n strings in Σ is divided into M buckets where each bucket contains N/M strings in Σ . (W.l.o.g., we assume that $M \mid N$.) If a client is to obtain m elements of the original database, he query (no more than) T elements from each of the M buckets. A batch code guarantees that based on the answers to the resulting $\leq M \cdot T$ queries, the client is able to efficiently reconstruct the m elements he was originally interested in.

When using a batch code, the storage provider only needs to store N elements. In addition, in the cryptographic scenario, the total computational complexity of the servers is $O(MT)$. Therefore, one is interested in simultaneously minimizing all three values N , M and T .

Several different batch codes were proposed in [6]. Batch codes have been recently studied very actively in the combinatorial setting. Namely, a *combinatorial batch code* (CBC) satisfies the additional requirement that every element of every bucket is equal to some element of the original database. (See for example [2], [1], [3].) New constructions of combinatorial batch codes, based on affine planes and transversal designs, were recently presented in [13].

CBCs suffer from some limitations. First, the requirement that each element in the bucket is equal to the element of the original database is generally not necessary. Relaxing this requirement can potentially lead to better parameter trade-offs. Second, batch codes are usually constructed using designs and related combinatorial structures. However, if such designs are employed in the practical settings, it might be difficult to come up with efficient retrieval algorithms for such codes due to the not-so-compact representation of the codes. As an alternative, we propose linear batch codes, where each bucket contains a linear combination of the elements in the original database. By using their flexible algebraic structure, better codes and more efficient retrieval algorithms can potentially be developed.

We stress that linear batch codes are also well suitable for the use in the *distributed data storage* [5]. The buckets can be viewed as servers. The reading of the requested data can be done “locally” from a small number of servers (for example, from those that are closer to the user, or connected by a higher-throughput links). The linear batch codes are generally fault-tolerant: if a small number of buckets stopped functioning, the

¹The work of the authors is supported in part by the research grants PUT405 and IUT2-1 from the Estonian Research Council and by the European Regional Development Fund through the Estonian Center of Excellence in Computer Science, EXCS. The work of V. Skachek is also supported in part by the EU COST Action IC1104.

data can be reproduced by reading data from (a small number) of other buckets. However, in order to achieve better *locality* and *repair bandwidth* (see [5]), linear batch codes with *sparse* generator matrices can be particularly useful.

In this paper, we develop a novel framework for analysis of linear batch codes, which is similar to that of classical error-correcting codes (ECCs). In particular, the encoding is represented by multiplying an information vector by an analog of a generator matrix. As we show, generator matrices of good binary linear batch codes are also generator matrices of good classical ECCs. This immediately gives us a set of tools and bounds from the classical coding theory for analyzing binary linear batch codes. The converse, however, is not true: not every good binary linear ECC is a good linear batch code. Then, we present a number of simple constructions of larger linear batch codes from the smaller ones. It worth mentioning that this novel framework opens a number of research directions related to linear batch codes. We conclude the paper with the list of some of the open questions.

The paper is structured as follows. The notations and some known results are presented in Section II. The properties of linear batch codes are analyzed in Section III. New construction methods of linear batch codes are presented in Section IV. The paper is summarized in Section V.

II. NOTATION AND KNOWN RESULTS

Let $[n] \triangleq \{1, 2, \dots, n\}$. We use notation $d_H(\mathbf{x}, \mathbf{y})$ to denote the Hamming distance between the vectors \mathbf{x} and \mathbf{y} , and notation $w_H(\mathbf{x})$ to denote the Hamming weight of \mathbf{x} . We also denote by $\mathbf{0}$ the row vector consisting of all zeros, and by \mathbf{e}_i the row vector having one at position i and zeros elsewhere (the length of vectors will be clear from the context). The binary entropy function is defined as $H_2(x) \triangleq -x \log_2 x - (1-x) \log_2 (1-x)$.

Definition 1 ([6]). *Let Σ be a finite alphabet. We say that \mathcal{C} is an $(n, N, m, M, t)_\Sigma$ batch code over a finite alphabet Σ if it encodes any string $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \Sigma^n$ into M strings (buckets) of total length N over Σ , namely $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_M$, such that for each m -tuple (batch) of (not necessarily distinct) indices $i_1, i_2, \dots, i_m \in [n]$, the entries $x_{i_1}, x_{i_2}, \dots, x_{i_m}$ can be retrieved by reading at most t symbols from each bucket. The ratio $R \triangleq n/N$ is called the rate of the code.*

If for the code \mathcal{C} it holds that $t = 1$, then we use notation $(n, N, m, M)_\Sigma$ for it. This corresponds to an important special case when only one symbol is read from each bucket.

Note that the buckets in this definition correspond to the devices in the above example, the encoding length N to the total storage, and the parameter t to the maximal load.

If $\Sigma = \mathbb{F}_q$ is a finite field, we also use notation $(n, N, m, M, t)_q$ (or $(n, N, m, M)_q$) to denote $(n, N, m, M, t)_\Sigma$ (or $(n, N, m, M)_\Sigma$, respectively).

Definition 2. *We say that an $(n, N, m, M, t)_q$ batch code is linear, if every entry of every bucket is a linear combination of original database elements.*

Before going further, we recall the following code composition from [6, Lemma 3.5].

Lemma 1 (Batch code composition). *Let \mathcal{C}_1 be an $(n_1, N_1 = M_1 n_2, m_1, M_1)_\Sigma$ batch code and \mathcal{C}_2 an $(n_2, N_2, m_2, M_2)_\Sigma$ batch code such that the length of each bucket in \mathcal{C}_1 is n_2 (in particular, $N_1 = M_1 n_2$). Then, there exists an $(n = n_1, N = M_1 N_2, m = m_1 m_2, M = M_1 M_2)_\Sigma$ batch code $\mathcal{C}_1 \otimes \mathcal{C}_2$.*

Thus, one can design batch codes by first considering special cases (like $M = N$), and then combining suitable batch codes to get rid of such restrictions.

III. LINEAR BATCH CODES

We start with the following example, which is based on so-called “subcube codes” in [6, Section 3.2].

Example 1. *Consider a database of n elements over \mathbb{F}_q , where the user wants to retrieve any m of them. Let t and n be integers, $2|n$. Divide the database $\mathbf{x} = (x_1, x_2, \dots, x_n)$ into two buckets, where bucket i , $i = 1, 2$, contains elements $(x_{(i-1)n/2+1}, x_{(i-1)n/2+2}, \dots, x_{in/2})$. The third bucket will contain elements $(x_1 + x_{n/2+1}, x_2 + x_{n/2+2}, \dots, x_{n/2} + x_n)$. This code is a linear $[n, N = 1.5n, m = 2t, M = 3, t]_q$ code for any $1 \leq t \leq n/2$. Observe, however, that the proposed code can be viewed as $n/2$ copies of the same $[2, 3, 2, 3, 1]_q$ subcube code.*

In what follows, we consider the case of a linear batch code \mathcal{C} with $t = 1$. Moreover, we limit ourselves to the case when $N = M$, which means that each encoded bucket contains just one element in \mathbb{F}_q .

Definition 3. *For simplicity we refer to a linear $(n, N = M, m, M)_q$ batch code as $[M, n, m]_q$ batch code.*

As before, let $\mathbf{x} = (x_1, x_2, \dots, x_n)$ be an information string, and let $\mathbf{y} = (y_1, y_2, \dots, y_M)$ be an encoding of \mathbf{x} . Due to linearity of the code, each encoded symbol y_i , $i \in [M]$, can be written as $y_i = \sum_{j=1}^n g_{j,i} x_j$ for some elements $g_{j,i} \in \mathbb{F}_q$, $j \in [n]$, $i \in [M]$. Then we can form the matrix \mathbf{G} as follows:

$$\mathbf{G} = \left(g_{j,i} \right)_{j \in [n], i \in [M]},$$

and thus

$$\mathbf{y} = \mathbf{x} \mathbf{G}.$$

The $n \times M$ binary matrix \mathbf{G} play a role similar to generator matrix for a classical linear ECC. In the sequel, we will call \mathbf{G} *generator matrix* of the batch code \mathcal{C} . We denote by \mathbf{G}_i the i -th row of \mathbf{G} and by $\mathbf{G}^{[i]}$ the i -th column of \mathbf{G} .

Observe that we can retrieve x_j from \mathbf{y} (for some $j \in [n]$) using $[M, n, m]_q$ batch code if there exists a linear combinations of columns in \mathbf{G} over \mathbb{F}_q , which is equal to \mathbf{e}_j . Moreover, the following generalization of this property holds.

Property 1. *Let \mathcal{C} be an $[M, n, m]_q$ batch code. It is possible to retrieve $x_{i_1}, x_{i_2}, \dots, x_{i_m}$ simultaneously if there exist m non-intersecting sets of indices of columns in \mathbf{G} , and*

for the r -th set there exists a linear combination of columns of \mathbf{G} indexed by that set, which equals to the column vector $\mathbf{e}_{i_r}^T$, for all $r \in [m]$.

Proof: Let

$$\mathbf{G} \triangleq [\mathbf{G}^{[1]} \mid \mathbf{G}^{[2]} \mid \dots \mid \mathbf{G}^{[M]}],$$

where $\mathbf{G}^{[\ell]}$ is the ℓ -th column in \mathbf{G} . Let T_1, T_2, \dots, T_m be non-intersecting sets of indices, such that for each $r \in [m]$

$$\mathbf{e}_{i_r}^T = \sum_{\ell \in T_r} \alpha_\ell \cdot \mathbf{G}^{[\ell]},$$

where all $\alpha_\ell \in \mathbb{F}_q$. Due to linearity, the encoding of $\mathbf{x} = (x_1, x_2, \dots, x_n)$ can be written as

$$\mathbf{y} = (y_1, y_2, \dots, y_M) = \mathbf{x} \cdot \mathbf{G}.$$

Then,

$$\begin{aligned} x_{i_r} &= \mathbf{x} \cdot \mathbf{e}_{i_r}^T \\ &= \mathbf{x} \cdot \left(\sum_{\ell \in T_r} \alpha_\ell \cdot \mathbf{G}^{[\ell]} \right) \\ &= \sum_{\ell \in T_r} \alpha_\ell (\mathbf{x} \cdot \mathbf{G}^{[\ell]}) \\ &= \sum_{\ell \in T_r} \alpha_\ell \cdot y_\ell, \end{aligned}$$

and therefore the value of x_{i_r} can be obtained by querring only the values of y_ℓ for $\ell \in T_r$. The conclusion follows from the fact that all T_r do not intersect. ■

In the rest of the paper we assume that the retrieving server performs only linear operations over the columns of the matrix \mathbf{G} (in other words, it only adds and subtracts y_1, y_2, \dots, y_M , and multiplies them by the elements in \mathbb{F}_q). This is a standard assumption in many areas of linear coding (in particular, in network and index coding). In that case, the condition in Property 1 becomes both necessary and sufficient.

Example 2. Consider the following linear binary batch code \mathcal{C} whose 4×9 generator matrix is given by

$$\mathbf{G} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

Let $\mathbf{x} = (x_1, x_2, x_3, x_4)$, $\mathbf{y} = \mathbf{x}\mathbf{G}$.

Assume that we want to retrieve the values of (x_1, x_1, x_2, x_2) . Consider, for example, the following

combinations of the columns of \mathbf{G} :

$$\begin{aligned} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} &= \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \\ \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \\ \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}. \end{aligned}$$

Then, we can retrieve (x_1, x_1, x_2, x_2) from the following set of equations:

$$\begin{cases} x_1 = y_1 \\ x_1 = y_2 + y_3 \\ x_2 = y_5 + y_8 \\ x_2 = y_4 + y_6 + y_7 + y_9 \end{cases}.$$

Moreover, it is straightforward to verify that any 4-tuple $(x_{i_1}, x_{i_2}, x_{i_3}, x_{i_4})$, where $i_1, i_2, i_3, i_4 \in [4]$, can be retrieved by using columns indexed by some four non-intersecting sets of indices in [9]. Therefore, the code \mathcal{C} is a $[9, 4, 4]_2$ batch code. As a matter of fact, this code is the two-layer construction of “subcube code” in [6, Section 3.2].

Next, we state the following simple lemmas.

Lemma 2. Let \mathcal{C} be an $[M, n, m]_q$ batch code. Then, each row of \mathbf{G} has Hamming weight at least m .

Proof. Consider row j , for an arbitrary $j \in [n]$. We can retrieve the combination (x_j, x_j, \dots, x_j) if there are m non-intersecting sets of columns, such that sum of the elements in each set is equal \mathbf{e}_j^T . Therefore, there are at least m columns in \mathbf{G} with a nonzero entry in position j . □

Lemma 3. Let \mathcal{C} be an $[M, n, m]_q$ batch code. Then, the matrix \mathbf{G} is a full rank matrix.

Proof. We should be able to recover any combination of size m of $\{x_1, x_2, \dots, x_n\}$. Then, the column vectors

$$\begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \dots, \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

are all in the column space of \mathbf{G} . Therefore, the column space of \mathbf{G} has dimension n , and so the matrix is full rank. □

The following theorem is the main result of this section. The presented proof of this theorem works only for *binary* batch codes. However, binary codes are very important special case of batch codes, as typical practical applications use binary representation of information. The proof uses the fact

that the codes are binary, — we are not aware of a simple generalization of this proof to nonbinary case.

Theorem 4. *Let \mathcal{C} be an $[M, n, m]_2$ batch code \mathcal{C} over \mathbb{F}_2 . Then, \mathbf{G} is a generator matrix of the classical error-correcting $[M, n, \geq m]_2$ code.*

Proof. Let \mathbb{C} be a classical ECC, whose generating matrix is \mathbf{G} . It is obvious that the length of \mathbb{C} is M . Moreover, since the matrix \mathbf{G} is a full rank matrix due to Lemma 3, we obtain that the dimension of \mathbb{C} is n . Thus, the only parameter in question is the minimum distance of \mathbb{C} .

In order to show that the minimum distance of \mathbb{C} is at least m , it will be sufficient to show that any non-zero linear combination of the rows of \mathbf{G} has Hamming weight at least m . Consider an arbitrary linear combination of the rows of \mathbf{G} , whose indices are given by a set $T \neq \emptyset$,

$$\mathbf{z} = \sum_{i \in T} \mathbf{G}_i.$$

Take an arbitrary index $i_0 \in T$. Due to the properties of the batch codes we should be able to recover $(x_{i_0}, x_{i_0}, \dots, x_{i_0})$ from \mathbf{y} . Therefore, there exist m disjoint sets of indices $S_1, S_2, \dots, S_m, S_i \subseteq [M]$, such that for all $i \in [m]$:

$$\sum_{j \in S_i} \mathbf{G}^{[j]} = \mathbf{e}_{i_0}^T. \quad (1)$$

Now, consider the sub-matrix \mathbf{M}_i of \mathbf{G} which is formed by the rows of \mathbf{G} indexed by T and the columns of \mathbf{G} indexed by S_i . Due to (1), the row of \mathbf{M}_i that corresponds to the row i_0 in \mathbf{G} , has an odd number of ones in it. All other rows of \mathbf{M}_i contain an even number of ones. Therefore, the matrix \mathbf{M}_i contains an odd number of ones. This means that the vector of \mathbf{z} will also contain an odd number of ones in the positions given by the set S_i . This odd number is at least one.

We conclude that \mathbf{z} contains at least one ‘1’ in positions given by the set S_i , for all $i \in [m]$. The sets S_i are disjoint, and therefore the Hamming weight of \mathbf{z} is at least m . \square

Example 3. *The converse of Theorem 4 is generally not true. In other words, if \mathbf{G} is a generator matrix of a classical error-correcting $[M, n, m]_2$ code, then the corresponding code \mathcal{C} is not necessarily an $[M, n, m]_2$ batch code. For example, take \mathbf{G} to be a generator matrix of the classical $[4, 3, 2]_2$ ECC as follows:*

$$\mathbf{G} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

Let $\mathbf{x} = (x_1, x_2, x_3)$, $\mathbf{y} = (y_1, y_2, y_3, y_4) = \mathbf{x}\mathbf{G}$.

It is impossible to retrieve (x_2, x_3) . This can be verified by the fact that

$$x_2 = y_1 + y_2 = y_3 + y_4 \quad \text{and} \quad x_3 = y_1 + y_3 = y_2 + y_4,$$

and so one of the y_i ’s is always needed to compute each of x_2 and x_3 .

Corollary. The topic of linear ECCs was very intensively studied over the years. Various well-studied properties of linear

ECCs, such as MacWilliams identities [9], apply also to linear batch codes due to Theorem 4 (for $t = 1$, $M = N$ and $q = 2$). A variety of bounds on the parameters of ECCs, such as sphere-packing bound (2), Plotkin bound (3), Griesmer bound (4), Elias-Bassalygo bound (5), McEliece-Rodemich-Rumsey-Welch bound (6) [11] (see also [12, Chapter 4], [10]) apply to the parameters of linear binary $[M, n, m]$ batch codes.

$$2^{M-n} \geq \sum_{i=0}^{\lfloor (m-1)/2 \rfloor} \binom{M}{i} \quad (2)$$

$$m \leq \frac{M \cdot 2^{n-1}}{2^n - 1} \quad (3)$$

$$M \geq \sum_{i=0}^{n-1} \left\lceil \frac{m}{2^i} \right\rceil \quad (4)$$

$$\frac{n}{M} \leq 1 - H_2 \left(\frac{1}{2} \left(1 - \sqrt{1 - 2 \frac{m}{M}} \right) \right) + o(1) \quad (5)$$

$$\frac{n}{M} \leq H_2 \left(\frac{1}{2} - \frac{\sqrt{m(M-m)}}{M} \right) + o(1) \quad (6)$$

IV. CONSTRUCTIONS OF NEW CODES

In this section we present several simple methods to construct new linear batch codes from the existing ones.

Theorem 5. *Let \mathcal{C}_1 be an $[M_1, n, m_1]_q$ batch code and \mathcal{C}_2 be an $[M_2, n, m_2]_q$ batch code. Then, there exists an $[M_1 + M_2, n, m_1 + m_2]_q$ batch code.*

Proof. Let \mathbf{G}_1 and \mathbf{G}_2 be $n \times M_1$ and $n \times M_2$ generator matrices corresponding to \mathcal{C}_1 and \mathcal{C}_2 , respectively. Consider the following $n \times (M_1 + M_2)$ matrix

$$\hat{\mathbf{G}} = [\mathbf{G}_1 \mid \mathbf{G}_2].$$

This matrix corresponds to a batch code of length $M_1 + M_2$ with n variables. It is sufficient to show that any combination of $m_1 + m_2$ variables can be retrieved. By the assumption, the first (any) m_1 variables can be retrieved from the first M_1 coordinates of \mathbf{y} and the last m_2 variables can be retrieved from the last M_2 coordinates of \mathbf{y} . This completes the proof. \square

Theorem 6. *Let \mathcal{C}_1 be an $[M_1, n_1, m_1]_q$ batch code and \mathcal{C}_2 be an $[M_2, n_2, m_2]_q$ batch code. Then, there exists an $[M_1 + M_2, n_1 + n_2, \min\{m_1, m_2\}]_q$ batch code.*

Proof. As before, denote by \mathbf{G}_1 and \mathbf{G}_2 the $n_1 \times M_1$ and $n_2 \times M_2$ generator matrices corresponding to \mathcal{C}_1 and \mathcal{C}_2 , respectively. Consider the following $(n_1 + n_2) \times (M_1 + M_2)$ matrix

$$\hat{\mathbf{G}} = \left[\begin{array}{c|c} \mathbf{G}_1 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{G}_2 \end{array} \right].$$

The matrix $\hat{\mathbf{G}}$ corresponds to a batch code of length $M_1 + M_2$ with $n_1 + n_2$ variables. Moreover, any combination of $\min\{m_1, m_2\}$ variables can be retrieved. If all unknowns are from $\{x_1, x_2, \dots, x_{n_1}\}$, then they can be retrieved by

using only the first M_1 columns of \hat{G} . If all unknowns are from $\{x_{n_1+1}, x_{n_1+2}, \dots, x_{n_1+n_2}\}$, then they can be retrieved by using only the last M_2 columns of \hat{G} . Generally, some unknowns can be retrieved by using combinations of the first M_1 columns, while the other unknowns are retrieved using combinations of the last M_2 columns. Since the number of unknowns is at most $\min\{m_1, m_2\}$, we can always retrieve all of them simultaneously. \square

The next theorem presents another construction of batch code from a smaller batch code.

Theorem 7. *Let \mathcal{C} be an $[M, n, m]_q$ batch code, and let G be the corresponding $n \times M$ matrix. Then, the code $\hat{\mathcal{C}}$, defined by the $(n+1) \times (M+m)$ matrix*

$$\hat{G} = \left(\begin{array}{c|cccc} & & & & 0 & 0 & \cdots & 0 \\ & & & & 0 & 0 & \cdots & 0 \\ & & & & \vdots & \vdots & \cdots & \vdots \\ & & & & 0 & 0 & \cdots & 0 \\ \hline \bullet & \bullet & \bullet & \cdots & \bullet & 1 & 1 & \cdots & 1 \end{array} \right)$$

$\underbrace{\hspace{10em}}_M \quad \underbrace{\hspace{10em}}_m$

is an $[M+m, n+1, m]$ batch code, where \bullet stands for an arbitrary element in \mathbb{F}_q .

Proof As before, let $\mathbf{x} = (x_1, x_2, \dots, x_n, x_{n+1})$ and $\mathbf{y} = (y_1, y_2, \dots, y_{M+m}) = \mathbf{x}\hat{G}$. Assume that we want to retrieve the vector $\mathbf{z} = (x_{i_1}, x_{i_2}, \dots, x_{i_m})$.

Take a particular x_{i_j} in \mathbf{z} , $j \in [m]$. Consider two cases. If $i_j \neq n+1$ then, since \mathcal{C} is a batch code, we have

$$x_{i_j} = \sum_{\ell \in T_{i_j}} y_\ell + \xi \cdot x_{n+1},$$

where $T_{i_j} \subseteq [M]$ and $\xi \in \mathbb{F}_q$. In that case, if $\xi = 0$, then $x_{i_j} = \sum_{\ell \in T_{i_j}} y_\ell$. If $\xi \neq 0$, then $x_{i_j} = \sum_{\ell \in T_{i_j}} y_\ell + \xi \cdot y_{M+j}$. Observe that all T_{i_j} are disjoint due to the properties of a batch code.

In the second case, $i_j = n+1$, and we simply set $x_{i_j} = x_{n+1} = y_{M+j}$.

In both cases, we used sets $\{y_\ell : \ell \in T_{i_j} \cup \{M+j\}\}$ to retrieve x_{i_j} . These sets are all disjoint for $j \in [m]$.

We conclude that all m unknowns x_{i_j} , $j \in [m]$, can be retrieved simultaneously. \square

V. DISCUSSION

In this paper, we studied linear batch codes. We defined generator matrices of such codes. We also showed that a generator matrix of a linear $[M, n, m]_2$ batch code is also a generator matrix of a classical $[M, n, m]_2$ ECC. The converse is not necessarily true. Finally, we presented several simple ways to construct new linear batch codes from smaller codes.

Since linear batch codes are closely related to linear ECCs, various well-known properties of linear ECCs, and in particular bounds on their parameters (2)-(6) apply also to linear binary batch codes (for $t = 1$ and $M = N$). Linear structure of batch codes can potentially be exploited in order to develop

efficient retrieval algorithms. Therefore, linear batch codes are natural candidates for practical applications, such as load balancing, CPIR and distributed storage. However, a lot of questions remain open. We list some of them below.

- 1) Can the connection between linear batch codes and ECCs be extended to nonbinary codes?
- 2) Construct linear batch codes with better trade-offs between their parameters.
- 3) Construct linear batch codes suitable for distributed storage settings, in particular codes having sparse generator matrices. Obtain bounds on locality and repair bandwidth for such codes.
- 4) Do linear batch codes have as good parameters as their nonlinear counterparts do?
- 5) Develop efficient retrieval algorithms for batch codes.

Answering some of these questions could help in developing of new and more efficient batch codes, which can potentially be used in practical applications.

ACKNOWLEDGEMENT

We thank Dominique Unruh for helpful discussions.

REFERENCES

- [1] S. Bhattacharya, S. Ruj, and B. Roy, Combinatorial batch codes: a lower bound and optimal constructions, *Advances in Mathematics of Communications*, vol. 6, no. 2, pp. 165–174, 2012.
- [2] R.A. Brualdi, K. Kiernan, S.A. Meyer, and M.W. Schroeder, Combinatorial batch codes and transversal matroids, *Advances in Mathematics of Communications*, vol. 4, no. 3, pp. 419–431, 2010.
- [3] C. Bujtás and Z. Tuza, Batch codes and their applications, *Electronic Notes in Discrete Mathematics*, vol. 38, pp. 201–206, 2011.
- [4] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, Private information retrieval, *Proc. 36th Symp. on Foundations of Comp. Science (FOCS)*, pp. 41–50, 1995.
- [5] A.G. Dimakis, P.B. Godfrey, Y. Wu, M.J. Wainwright, and K. Ramchandran, Network coding for distributed storage systems, *IEEE Trans. Inform. Theory*, vol. 59, no. 9, pp. 4539–4551, 2010.
- [6] Y. Ishai, E. Kushilevitz, R. Ostrovsky, and A. Sahai, Batch codes and their applications, *Proc. of the 36th ACM Symposium on Theory of Computing (STOC)*, June 2004, Chicago, IL.
- [7] E. Kushilevitz and R. Ostrovsky, Replication is NOT needed: SINGLE database, computationally-private information retrieval, *Proc. 38th Symp. on Foundations of Comp. Science (FOCS)*, pp. 364–373, 1997.
- [8] H. Lipmaa, First CPIR protocol with data-dependent computation, *Proc. International Conference on Information Security and Cryptology (ICISC)*, pp. 193–210, 2009.
- [9] F.J. MacWilliams, A theorem on the distribution of weights in a systematic code, *Bell System Tech. J.*, 42 (1963), 79–94.
- [10] F. J. MacWilliams and N. J. A. Sloane, “The Theory of Error-Correcting Codes,” Amsterdam, The Netherlands; North-Holland, 1978.
- [11] R.J. McEliece, E.R. Rodemich, H. Rumsey, and L.R. Welch New upper bounds on the rate of a code via the Delsarte-MacWilliams inequalities, *IEEE Trans. Inform. Theory*, vol. IT-23, pp. 157–166, Mar. 1997.
- [12] R. M. Roth, “Introduction to Coding Theory,” Cambridge University Press, Cambridge, United Kingdom, 2006.
- [13] N. Silberstein and A. Gál, Optimal Combinatorial Batch Codes based on Block Designs, preprint is available online <http://arxiv.org/abs/1312.5505>.
- [14] D. Stinson, R. Wei, and M. Paterson, Combinatorial batch codes, *Advances in Mathematics of Communications*, vol. 3, no. 1, pp. 13–17, 2009.